Code Generation: Results Achieved with Zero-shot Prompting When Not Asking for a Rationale

LLM	Boolean judgment		No Rationale	
	#	%	#	%
DeepSeek Coder 1.3B	687	42.54%	421	26.07%
DeepSeek Coder 6.7B	22	1.36%	0	0.00%
DeepSeek Coder 33B	4	0.25%	3	0.19%
CodeLlama 7B	1	0.06%	1	0.06%
CodeLlama 13B	3	0.19%	8	0.50%
CodeLlama 34B	2	0.12%	3	0.19%
GPT-3.5-turbo	0	0.00%	1	0.06%
GPT-4-turbo	27	1.67%	0	0.00%

Table 1: Number (#) and percentage (%) of instances for which the LLMs did not manage to output a valid judgment. Comparison between the scenarios where asking for a rationale (left) or and not asking for a rationale (right).

DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
0.05	0.02	0.15	-0.06	0.07	0.23	0.09	0.19

Table 2: Code Generation: Kappa agreement between the binary judgments of the LLMs and the pass/fail (0 or 1) ground truth.

	DSC 1.3B	DSC 6.7B	DSC 33B	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
DSC 1.3B	0.60	0.54	0.45	0.50	0.47	0.53	0.57	0.55	-0.12	(N)	(N)	*** (L)
DSC 6.7B	0.74	0.70	0.62	0.67	0.63	0.65	0.57	0.62	-0.01	(N)	(N)	*** (L)
DSC 33B	0.34	0.39	0.31	0.34	0.32	0.34	0.32	0.40	-0.60	(N)	(N)	*** (L)
CL 7B	-0.16	-0.21	-0.27	-0.21	-0.28	-0.22	-0.12	-0.21	-0.93	(N)	(N)	*** (L)
CL 13B	0.59	0.53	0.43	0.54	0.46	0.47	0.59	0.56	-0.10	(N)	(N)	*** (L)
CL 34B	0.24	0.23	0.13	0.17	0.17	0.12	0.28	0.30	-0.52	* (N)	* (N)	*** (L)
GPT-3.5	0.50	0.55	0.47	0.51	0.51	0.52	0.52	0.53	-0.23	(N)	(N)	*** (L)
GPT-4	0.35	0.42	0.35	0.34	0.35	0.32	0.42	0.51	-0.42	*** (N)	*** (N)	*** (L)
Average (all)	**0.40**	**0.40**	**0.31**	**0.36**	**0.33**	**0.34**	**0.39**	**0.41**	**-0.37**	-	-	-
Average (large)	**0.40**	**0.43**	**0.34**	**0.38**	**0.36**	**0.35**	**0.42**	**0.46**	**-0.37**	**-**	**-**	**-**

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 3: Average of differences between the LLM judgments (0 or 1) and the ground truth (*i.e.,* 1 if the method passes the test and 0 otherwise). Last three columns report adj. p-value and effect size when comparing the judgements each LLM gave to functions it generated against those it gave when judging functions (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.

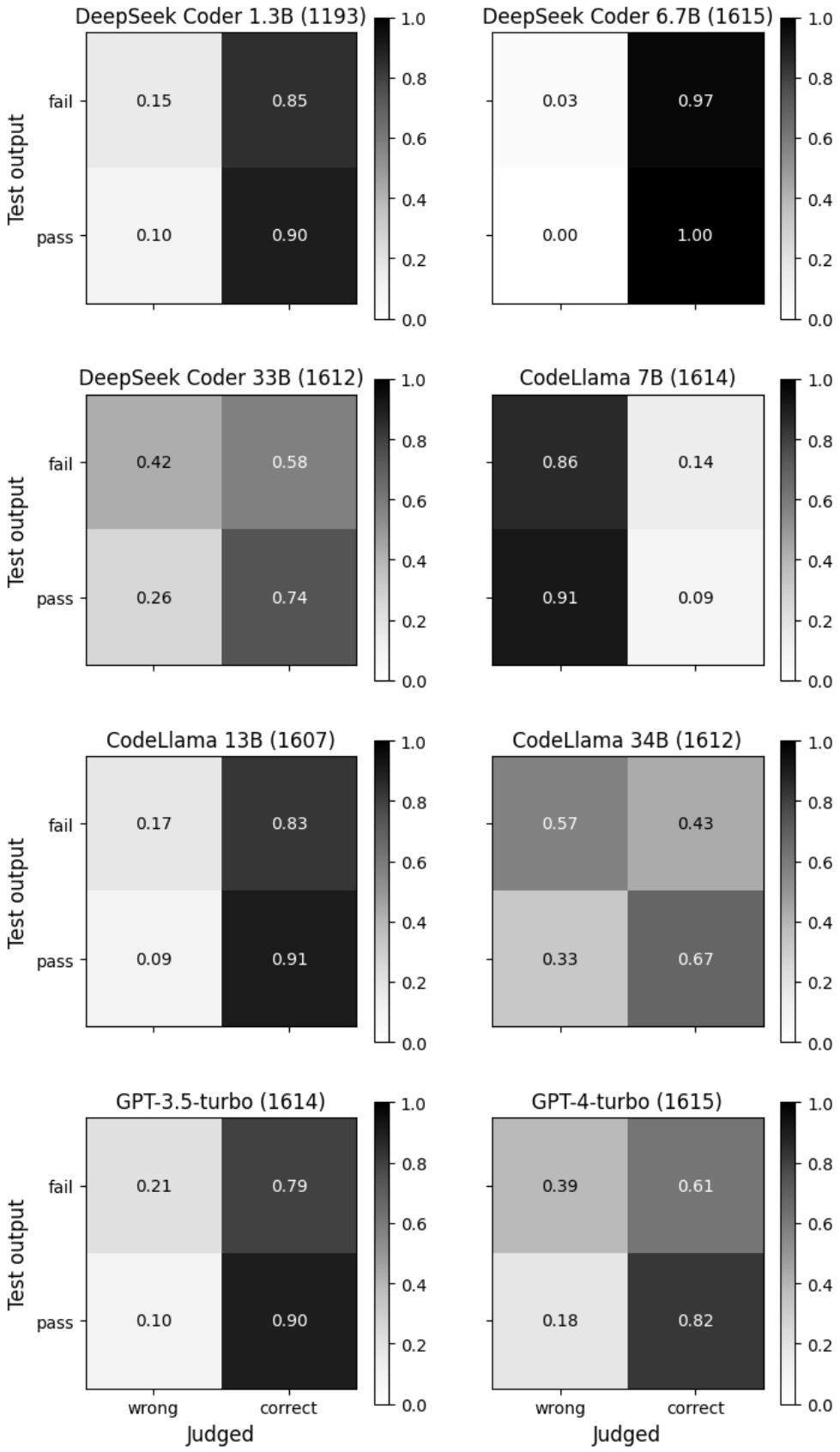

Figure 1: Code Generation: Code Generation: Confusion matrices for LLM's judgment.